DATA CLUSTERING USING MIN-MIN ROUGHNESS AND ITS APPLICATION TO CLUSTER PATIENTS SUSPECTED DIABETICS

MOHD RIDZUAN BIN BAHARIN

A report submitted in partial fulfillment
of the requirements for the award
of the degree of
Bachelor of Computer Science (Software Engineering)

Faculty of Computer System & Software Engineering
Universiti Malaysia Pahang

JUNE, 2012

ABSTRACT

In the context of information technology nowadays, there are many data exists. All of this data are scrambled over inside the computer and with the presence of internet, even more data exist. The problem with this is, when we want the needed data only, there are too many to look for and they are all scrambled over the internet databases. Therefore, there are techniques that are proposed that will provide a way to automatically mine the data and obtain only meaningful data from the huge data over the internet. The area discussed in this research is Knowledge Discovery in Databases (KDD) and the technique used is Minimum-Minimum Roughness (MMR). The dataset used will be the dataset of diabetic patients. By using this MMR technique, I intended to cluster the diabetic dataset n which each cluster will contain the data most related to each other.

ABSTRAK

Dengan adanya teknologi informasi sekarang ini, jumlah data-data semakin banyak. Data-data ini tersimpan di dalam computer-komputer dan dengan kehadiran internet, lebih banyak lagi data yandg ada. Perkara ini menimbulkan masalah apabila kita inginkan data-data yang kita mahukan, tetapi data-data yg ada adalah terlalu banyak dan berselerak di internet. Oleh sebab itu, terdapat teknik-teknik yang diperkenalkan untuk mengatas masalah ini. bidang yang dibincangkan ialah bidang Knowledge Discovery in Databases dan teknik yang digunakan ialah teknik Minimum-Minimum Roughness. Set data yang digunakan ialah set data pesakit-pesakit diabetis. Dengan menggunakan teknik MMR ini, sy bertujuan untuk mengklusterkan data tersebut dimana setiap kluster mengandungi data-data yang berkaitan denagn satu sama lain.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
1	**INTRODUCTION**	1
	1.1 Background	1
	1.2 Problem Statement	3
	1.3 Scopes	4
	1.4 Objectives	4
	1.5 Thesis Organization	4
2	**LITERATURE REVIEW**	5
	2.1 Diabetes	5
	2.1.1 Diabetes Description	5
	2.1.2 Diabetes Symptoms	6
	2.1.3 Diabetes in the World	7
	2.1.4 Diabetes in Asia	8
	2.1. 5 Diabetes in Malaysis	9
	2.1.6 Patients Suspected Diabetes	11
	2.2 Knowledge Discovery in Databases	12
	2.2.1 Definition of KDD	12
	2.2.2 KDD Process	13
	2.2.3 Examples of KDD Processes	15
	2.2.4 Application of KDD in computer science fields	15
	2.3. Data Mining	16

2.3.1. Definition of DM 16
2.3.2. Examples of DM 18
2.3.3. Applications of DM in computer science fields 19
2.4. Data Clustering 20
2.4.1. Definition 20
2.4.2. Classification vs Clustering 21
2.4.3. Clustering Techniques 23
2.4.4. Clustering on Numerical Dataset 24
2.4.5. Clustering on Categorical Dataset 25
2.4.6. Applications of Clustering Techniques 26
2.5. Rough Set Theory 27
2.5.1. Rough set 28
2.5.2. Fuzzy set 29
2.5.3. Relation between fuzzy and rough set theories 29
2.5.4. Applications of rough set 30
2.6. Rough Clustering 31
2.6.1. Application of rough set in data clustering 32
2.6.2. Rough set theory in categorical data clustering 32

3 METHODOLOGY 34
3.1. Rough Set Theory 34
3.1.1. Information System 35
3.1.2. Indiscernibility Relation 39
3.1.3. Approximation Space 40
3.1.4. Set Approximations 40
3.2. Min-Min Roughness 45

3.2.1. Selecting a clustering attribute 45

3.2.2. Model for selecting a clustering attribute 46

3.2.3. Min-Min Roughness Technique 46

3.2.4. Algorithm 47

3.2.5. Example 49

3.3. Object Splitting model 91

3.3.2. The splitting point attributes a4 is determined 92

3.3.3. Cluster Purity 92

4 EXPECTED RESULTS AND DISCUSSION 93

4.1. Datasets 93

4.1.1. Small datasets 93

4.1.2. Large dataset (real world dataset) 94

4.1.3. Benchmark dataset 94

4.2. MMR Software Development 94

4.2.1. Interface 95

5 CONCLUSION 97

REFERENCE 98

LIST OF TABLES

TABLE NO.	TITLE	PAGE
3.1	An information system	36
3.2	A diabetic decision system	37
3.3	Step-by-step Max-Max Roughness	46
3.4	An information system in MMR	49
3.5	Mean roughness a_1	85
3.6	Mean roughness a_2	86
3.7	Mean roughness a_3	86
3.8	Mean roughness a_4	87
3.9	Mean roughness a_5	87
3.10	Mean roughness a_6	88
3.11	Mean roughness a_7	88
3.12	Mean roughness a_8	88
3.13	Mean roughness a_9	89
3.14	Mean roughness a_{10}	90
3.15	Minimum roughness	91
3.16	MMR value	91

LIST OF FIGURES

TABLE NO.	TITLE	PAGE
1.1	KDD Process	2
2.1	Prevalence of diabetes by age group in Malaysia	10
2.2	Prevalence of diabetes in Malaysia by states	10
2.3	Overview of the steps that compose the KDD process	13
2.4	Classification	21
2.5	Clustering	22
2.6	Clustering process	23
3.1	Set approximations	42
3.2	Model for selecting a clustering attribute	46
3.3	Result of clustering	92
4.1	Start Interface	95
4.2	Calculation Interface	95

CHAPTER I

INTRODUCTION

This chapter briefly discuss on the overview of this research. It contains six parts. The first part is introduction; follow by the problem statement. Next is the motivation, followed by the scopes of the research. After that are the objectives where the research's goal is determined and lastly is the thesis organization which briefly describes the structure of this thesis.

1.1 Background

Knowledge Discovery from Databases (KDD) is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. As a branch of machine learning, KDD encompasses a number of automated methods whereby useful information is mined from data stored in databases.

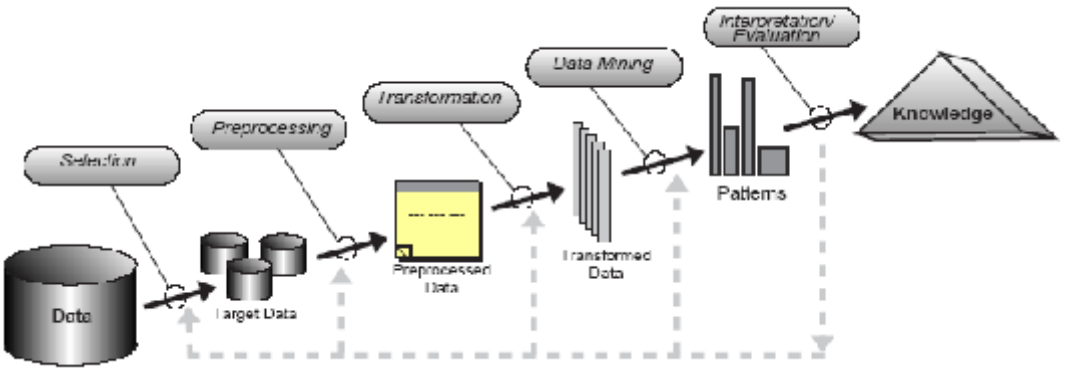

Figure 1.1: KDD Process

Data Mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data . Classification and Clustering are two of the Data Mining methods. Classification involves learning a function that maps (or classifies) a data item into one of several predefined classes, while clustering involve identifying a finite set of categories or clusters to describe the data.

Diabetes is a disease in which a person has high levels of sugar in the blood. It arises when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin produced. Insulin is needed to control blood sugar. Diabetes is a chronic, potentially debilitating and often fatal disease. It is one of the oldest known diseases; it is mentioned in Egyptian manuscript from around 1550 BCE, also indentified by an Indian physician Sushruta in 6th century BCE. Diabetes appears to have been a death sentence in the ancient era and even until now, there is no cure for diabetes. Globally, diabetes ranked as the fourth leading cause of death, in terms of disease. An estimated 3.8 million people die from diabetes-related causes each year. Such causes are: heart disease, stroke, kidney disease, nerve disease, diabetic eye disease others.

The United Nations estimates the number of people globally affected by diabetes to be 246 million and approximately half of those are in India, China, Nepal and other Asian countries.

In Malaysia, diabetes is a growing concern. In 1986, a survey, namely National Health and Morbidity Survey or NHMS for short, included diabetes as a major component in

the survey. Then in the second survey, the prevalence of diabetes in Malaysia was found to be 8.2%. There was an increase in prevalence as compared to the NHMS in 1986, which only reported 6.3% in Peninsular Malaysia. Universiti Kebangsaan Malaysia's Emeritus Professor Datuk Dr Khalid Abdul Kadir said there was a "diabetic explosion" in Malaysia and wondered whether enough was being done to stop it. He said an example could be seen among the Malays in Tanjung Karang, Selangor. The prevalence was four per cent in 1984 and 6.5 per cent in 1990. Two years ago, it shot up to 20 per cent. Dr Khalid said one in seven adults in Malaysia was a diabetic. Dr Khalid attributed the growing number of diabetic cases to the lack of physical activity and excess calories accumulation as one ages. "As the population ages, we are going to see more people with diabetes," he said, adding that diabetes, hypertension and obesity seldom killed a person but they contributed to heart diseases.

Despite of the fact that diabetes is caused by both lifestyle and genetic, the lifestyle factor, or in other word; diet, contribute much to diabetes. Based on a survey made in 1996, 16.6% of adult Malaysians are facing overweight problems while 4.4% of adult Malaysians are obese. Then another survey made in 2006, shows that 29.1 Malaysians are facing overweight problem while 14% Malaysian are obese. This increase in the number of Malaysians facing overweight problems and obesity will increase the number of chronic patients as 90% of the overweight and obese are facing diabetes.

1.2 Problem Statement

There are many data clustering methods that exist, however, most of them only dealt with only numerical data type. The problem is nowadays, in real life we are dealing with categorical data type which is multi-valued data. Also, there are uncertainties in these data that need to be handled.

The clustering of this diabetics' data involves multi-valued data. Therefore, rough set technique is used because it can handle the uncertainty and also deal with the multi-valued data of the diabetics.

Having the 8.2% prevalence of diabetes found in Malaysia, the data of those diabetics should be grouped into a balanced and meaningful cluster, and in this research, based on the symptoms of the diabetics. This grouping can help in classification of the diabetics and further investigation on that disease in Malaysia.

1.3 Scopes

The scopes of this research are:

i. The data used are from the diabetics of Hospital Ampuan Afzan.
ii. The clustering uses min-min roughness technique.
iii. Apply to diabetics dataset.

1.4 Objectives

There are a few objectives of this research:

i. To partition the patients in a meaningful way based on the symptoms' closeness.
ii. To apply the rough set clustering technique into a real life case.

1.5 Thesis Organization

The rest of this paper is organized as follows. Chapter II describes the notion of rough set. Chapter III describes the theory of rough set. Chapter IV describes the dataset, modeling process and min-min roughness data clustering. Chapter V describes the results from an application of rough set theory for clustering data and grouping diabetes patients following by discussion. Finally, the conclusion of this work is described in section 6.

CHAPTER II

LITERATURE REVIEW

This chapter briefly discusses on existing literature related with the proposed project. There are four main sections in this chapter. The first section introduces on the topic of diabetics. The second section describes some brief information on Knowledge Discovery in Databases (KDD). The third section describes Data Mining (DM) concept. The fourth section describes Data Clustering and finally a brief review of Rough Set Theory (RST) is described in the last section.

2.1. Diabetes

This section firstly presents a description and symptoms of diabetics. Further, information of diabetics in the world, Asia and Malaysia also presented. Finally, the last sub-section presents information of patient having pre-diabetes.

2.1.1. Diabetes Description

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Diabetic,

or a person with diabetes has a condition in which the quantity if glucose in the blood is too elevated (hyperglycemia), resulting in too much glucose building up in the blood which will eventually passes out of the body in urine. Glucose in the blood gives the energy to perform daily activities while insulin allows it the glucose to move from the blood into the liver, muscle, and fat cells, providing the essential energy and growth requirements. In diabetes, glucose in the blood cannot move into cells, so it stays in the blood, which harms the cells that need the glucose for energy and also harms certain organs and tissues exposed to the high level glucose. There are three types of diabetes; type 1 diabetes or insulin-dependent diabetes, is an auto-immune disease where the body is immune system destroys the insulin-producing beta cells in the pancreas; type 2 diabetes or non-insulin dependent diabetes, the most common form of diabetes, is characterized by insulin resistance and relative insulin deficiency, strongly genetic in origin but lifestyle factors such as excess weight, inactivity, high blood pressure and poor diet are major risk factors for its development; and gestational diabetes mellitus (GDM) or carbohydrate intolerance, usually developed during pregnancy and usually the carbohydrate intolerance returns to normal after the birth but the mother has a significant risk of developing permanent diabetes while the baby is more likely to develop obesity and impaired glucose tolerance and/or diabetes later in life.

2.1.2. Diabetes Symptoms

These are some of the common early warning sign of diabetes: excessive thirst that is unrelated to exercise, hot weather or short-term illness; excessive hunger, even after eaten; frequent urination; fatigue or tiredness possibly severe enough to make you fall asleep unexpectedly after meals; and sudden weight loss or dramatic change in weight. While many of the signs and symptoms of diabetes can also be related to other causes, testing for diabetes is very easy, and the constant/regular presence of these symptoms over an extended period of time should be cause to visit a doctor. Also, there are minor and less recognizable symptoms of diabetes, which are; blurry vision, occur because diabetes can lead to macular degeneration and eventual blindness; numbness, or tingling in the hands and feet may occur due to peripheral neuropathy, a symptom of diabetes,

causes nerve damage in the extremities; slow healing wounds, result of diabetes-related impaired immune system function; recurrent or hard-to-treat yeast infections in woman, another sign of impaired immune function; and dry or itchy skin, may result from peripheral neuropathy which affects circulation and proper sweat gland function.

2.1.3. Diabetes in the world

According to the report of diabetes published by WHO (World Health Organization) in 2009, there were at least 220 million diabetics, and the WHO also estimated that there will be at least 336 million diabetics in the world in 2030. There are 5% of global deaths which are caused from diabetes compilations. The WHO had stated a few facts about diabetes:

i. There are more than 346 million people worldwide have diabetes. There is an emerging global epidemic of diabetes that can be traced back to rapid increases in overweight, obesity and physical inactivity.
ii. Diabetes is predicted to become the seventh leading cause of death in the world by the year 2030. Total deaths from diabetes are projected to rise by more than 50% in the next 10 years.
iii. Type 2 diabetes is much more common than type 1 diabetes. Type 2 accounts for around 90% of all diabetes worldwide. Reports of type 2 diabetes in children, which previously rare, have increased worldwide. In some countries, it accounts for almost half of newly diagnosed cases in children and adolescents.
iv. Cardiovascular disease is responsible for between 50% and 80% of deaths in people with diabetes. Diabetes has become one of the major causes of premature illness and death in most countries, mainly through the increased risk of cardiovascular disease (CVD).
v. In 2004, an estimated 3,4 million people died from consequences of high blood sugar.

vi. 80% of diabetes deaths occur in low and middle income countries. In developed countries most people with diabetes are above the age of retirement, whereas in developing countries the most frequently affected are aged between 35 and 64.

vii. Diabetes is a leading cause of blindness, amputation and kidney failure due to lack of awareness about diabetes, combined with insufficient access to health services and essential medicines.

2.1.4. Diabetes in Asia

Research published in the medical journal Lancet reveals that life-threatening diabetes is becoming an epidemic not only in North America, but in Asia as well and it appears to be only getting worse. According to doctors at the Catholic University of Korea in Seoul, 194 million Asians were diabetic in 2003, a statistic that could soar to 330 million by the year 2025. The Lancet research suggests Asians are developing diabetes at younger age and at lower weight, they suffer longer complications and they also die earlier than people in developed countries. This onset of adult diabetes in increasingly younger populations will negatively affect Asian countries economically, as a result of higher health costs and mortality rates. According to the International Diabetes Federation (IDF)'s 2003 statistics, the top 5 countries with the largest number of diabetics were: India with 35.5 million diabetics; China with 23.8 million diabetics; USA with 16.0 million diabetics; Russia with 9.7 million diabetics; and Japan with 6.7 million diabetics. Also, there are four more countries in Asia among the top ten countries with highest number of diabetics; Indonesia, Pakistan, Bangladesh, and Philippines [9]. The WHO and the IDF predict that the number of diabetics on Asia could increase to 160 million by year 2025. Conservative estimates based on population growth and ageing and rate of urbanization in Asia show that India and China will remain the two countries with the highest numbers of people with diabetes by 2030. The current estimated age-adjusted prevalence of diabetes in China (4.2%) is expected to increase by 1.9% to 5.0% in 2030. However, results from a survey conducted in 2007-08 in China reported a higher prevalence (9.7%, including previously undiagnosed

diabetes). A similar increase is expected in India (from 7.8% in 2010 to 9.3 in 2030). Southern and Eastern Asian countries account for half the deaths attributable to diabetes worldwide and the consequences of a rising incidence and prevalence of diabetes and other chronic diseases will be particularly important, both locally and globally. A concurrent increase in the prevalence of overweight and obesity over the same time period has been observed in many Asian countries. In rural China, the prevalence of overweight has increased from 5.3% in men and 9.8% in women in 1992 to 13.6% in men and 14.4% in women n 2002. Corresponding figures for obesity were 0.5% in men and 0.7 % in women in 1992 and 1.8% in men and 3.0% in women in 2003. This trend has profound implications for the expected number of diabetes patients who will be diagnosed in this region for future decades. Type 2 is due primarily to lifestyle factors and genetics and most of Asian patients have a first-degree relative with diabetes. Also, most of the loci originally associated with diabetes in European populations have been replicated in Asian population. The urbanization and migration of Asians, which expected to increase, would be the cause of the rise in the global prevalence of diabetes.

2.1.5. Diabetes in Malaysia

Studies shows that the prevalence of type 2 diabetes is escalating at phenomenal scale and very likely we are heading towards epidemic proportions. Malaysia is too to be affected by this as there are major shift in the lifestyles and longetivity of the population. In other words, Malaysia has the right ingredients to set the scene for the explosion of diabetes [m]. Professor Datuk Dr Khalid Abdul Kadir, University Kebangsaan Malaysia Emeritus and also a professor of medicine at Monash University, said there was a "diabetic explosion" in Malaysia and wondered whether enough was being done to stop it. He said one in seven adults in Malaysia was a diabetic and more people below the age of 45 are getting diabetes. He also said that with modernization and economic progress, there would be an explosion of "metabolic catastrophe" in Asia, including Malaysia, due to obesity, hypertension and diabetes. According to him, in 1990, the prevalence of obesity and diabetes among the Orang Asli, the hunter-gatherers in the jungle fringes of Pahang or in settlements at Carey Islands and Ulu Langat outside Kuala Lumpur was

zero. But over five years, the Institute for Medical Research found 5% of the resettled Orang Asli had diabetes. He attributes the growing number of diabetic cases to the lack of physical activity and excess calories accumulation as one ages . Statistics pointed that Malaysia had the fourth highest number of diabetes cases in Asia, with 800, 000 in 2007. In the Malaysian Burden of Disease and Injury Study, it was estimated that for year 2000, there were 2,261 deaths attributed to diabetes in which 847 of them are men and 1404 are women. The earliest diabetes studies carried out in Malaysia were in 1960 and in 1966. The first National Health and Morbidity Survey (NHMS) in Malaysia was carried out in 1986 where prevalence of diabetes among adults of age 35 years old and above was found to 6.3%. Then 10 years later, in NHMS II, the figure had increased by one third to 8.3% among adults of age 30 years and above. This shocking rise spurred the initiation of numerous national healthy lifestyle campaigns by the Ministry of Health Malaysia. A national steering committee was set up to improve the screening and management of diabetes in primary and secondary care clinics. In 2006, the third NHMS is conducted. The result shows that diabetes is also detected in the younger age group, between the ages of 18 to 30 years old. Also, there was a general increasing trend in diabetes prevalence with age; from 2.0% in the 18-19 years old age group to a prevalence ranging between 20.8 to 26.2% among the 60-64 years old shown in Figure 1.

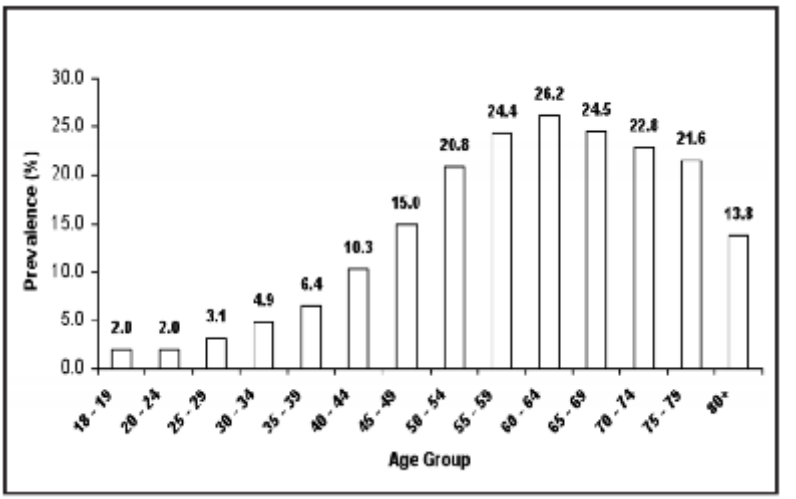

Figure 2.1: Prevalence of diabetes by age group in Malaysia

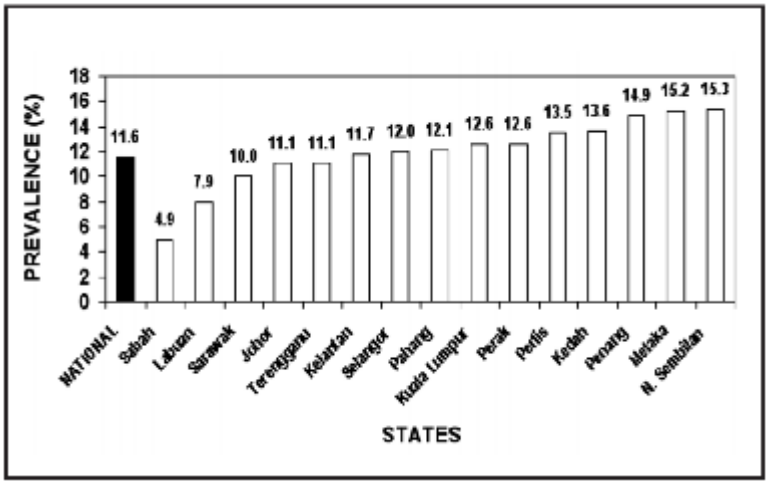

Figure 2.2: Prevalence of diabetes in Malaysia by states

Among the states, Negeri Sembilan, Malacca and Penang had the highest prevalence of diabetes at 15.3%, 15.2% and 14.9% respectively as shown in Figure 2. The prevalence was higher in the urban at 12.2% compared to the rural areas at 10.6%. No significant gender difference was observed; 11.9% in the males while 11.3% in the females.

2.1.6. Patient suspected Diabetes

Patient suspected diabetes can be also said as people having pre-diabetes or people with the risk of diabetes. Pre-diabetes is the state that occurs when a person's blood glucose levels are higher than normal but not high enough for a diagnosis of diabetes. This means that this people have the risk of getting diabetes. According to the American diabetes association, in the Diabetes Prevention Program, 11% of people with pre-diabetes developed type 2 diabetes each year during the average 3 years of follow up. Other studies show that many people with pre-diabetes develop type 2 diabetes in 10 years. People suspected diabetes might not know that they are having diabetes risk. In fact, many people that have diabetes do not realize it because of the symptoms develop so gradually, people often do not recognize them. Some people have no symptom at all.

There are many risk factors for type 2 diabetes, as shown below:

i. Obesity; the National Center for Health Statistics states that 30% of adults are obese. That is 60 million people. Greater weight means a higher risk of insulin resistance, because fat interferes with the body's ability to use insulin.

ii. Sedentary lifestyle; muscle cells have more insulin receptor than fat cells, so a person can decrease insulin resistance by exercising. Being more active also lowers blood sugar levels by helping insulin to be more effective.

iii. Unhealthy eating habits; 90% of people who have been diagnosed with type 2 diabetes are overweight. Unhealthy eating contributes largely to obesity. Too much fat, not enough fiber and too many simple carbohydrates all contribute to a diagnosis of diabetes.

iv. Family history and Genetics; it appears that people who have family members who have been diagnosed with type 2 diabetes are at a greater risk for developing it themselves. Lifestyle plays in important part n determining who gets diabetes.

v. Increased age; the older we get, the higher our risk of type 2 diabetes.

vi. High blood pressure and high cholesterol; having metabolic syndrome increases the risk of heart disease, stroke and diabetes.

vii. History of gestational diabetes; gestational diabetes affects 4% of all pregnant women. Many women who have gestational diabetes develop type 2 diabetes years later. Their babies are also at some risk for developing diabetes in later in life.

2.2. Knowledge Discovery in Databases

This section presents a definition, processes, examples and applications of Knowledge Discovery in Databases (KDD).

2.2.1. Definition of KDD

The rapidly growing volume and complexity of modern databases make the need for technology to describe and summarize the information they contain increasingly important. Knowledge Discovery in Databases (KDD) and data mining are new research areas that try to deal with this problem. KDD is defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. There are also many other terms, appearing in some articles and documents, carrying a similar or slightly different meaning, such as knowledge mining from database, knowledge extraction, data archeology, data grudging, data analysis and so on. The goal of KDD is to make the patterns of data understandable to humans. The data are a set of facts and pattern is an expression in some language describing a subset of the data or a model applicable to the subset. The discovered patterns should be valid on new data with some degree of certainty. The patterns need to be novel and potentially useful, that is lead to some benefit to the user or task. Having KDD means that the interesting knowledge, regularities or high level information can be extracted from the relevant sets of data in databases and be investigated from different angles, and large database thereby serves as rich and reliable sources for knowledge generation and verification.

2.2.2. KDD Processes

The KDD process involves using the database along with any required selection, preprocessing, subsampling and transformations of it; applying data mining methods (algorithms) to enumerate patterns from it; and evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge. The overall KDD process includes the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge.

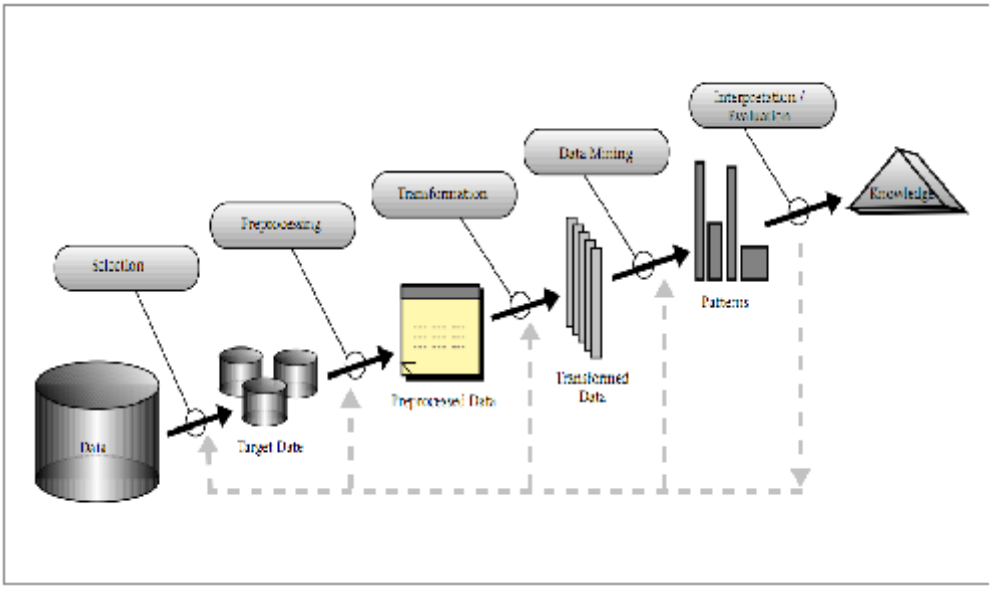

Figure 2.3: Overview of the steps that compose the KDD process

The process of KDD consists of the following steps:

i. Developing an understanding of the application domain, the relevant prior knowledge, and the goal(s) of the end user.

ii. Creating or selecting a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.

iii. Data cleaning and preprocessing: this step includes, removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.

iv. Data reduction and projection: finding useful features to represent the data depending in the goal of the task. This may include dimensionality reduction or transformation to reduce the effective number of variables under consideration or to find the invariant representations of the data.

v. Matching the goals to a particular data mining method such as summarization, classification, regression, clustering etc. Model and

hypothesis selection, choosing the data mining algorithm(s) and methods to be used for searching for data patterns.

vi. Exploratory analysis and model and hypothesis selection: choosing the data mining algorithms(s) and selecting method(s) to be used for searching for the data patterns. This process includes deciding which models and parameters might be appropriate and matching particular data mining method with the overall criteria of the KDD process.

vii. Data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data mining method by correctly performing the preceding steps.

viii. Interpreting mined patterns: possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.

ix. Acting on discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

2.2.3. Examples of KDD Processes

One example of a system that applies KDD in astronomy area is SKICAT, a system used by astronomers to perform image analysis, classification and cataloging of sky objects from sky-survey images. In its first application, the system was used to process 3 terabytes (10^{12} bytes) of image data resulting from the Second Palomar Observatory Sky Survey, where it is estimated that on the order of 10^{9} sky objects are detectable.

SKICAT can outperform humans and traditional computational techniques in classifying faint sky objects.

While in marketing, the primary application is database marketing systems, which analyze customer databases to identify different customer groups and forecast their behavior. Another notable marketing application is market-basket analysis system, which find patterns such as, “If customer bought X, he/she is also likely to buy Y and Z”. Such patterns are valuable to retailers.

KDD is also applied in fraud detection system such as HNC Falcon and Nestor PRISM systems which are used for monitoring credit card fraud, watching over millions of accounts. Also, the FAIS system from the U.S. Treasury Financial Crimes Enforcement Network, is used to identify financial transactions that might indicate money laundering activity.

Lastly, a novel and increasingly important type of discovery is one based on the use of intelligent agents to navigate through an information-rich environment. These systems ask the user to specify a profile of interest and search for related information among a wide variety of public domain and proprietary sources.

2.2.4. Application of KDD in computer science fields

Computer science is the study of the theoretical foundations of information and computation and of practical techniques for their implementation and application in computer system. In today’s society, the Information Technology (IT) is an increasingly part of all economic, technological, educational and even cultural sectors. Through applications such as e-commerce, networking, and digital administration, the IT evolution has become one of the most important factors in shaping the future of our social system. Currently, many kinds of data are being generated and stored about all kind of human endeavors for example like the widespread use of the bar codes for most commercial products, the computerization of many business and government transaction, and the advances in data collection tools have provided us with huge amounts of data. These data are stored or recorded in the form of computer databases, where the computer technology can easily access it. Millions of databases have been